

**Indian Institute of Technology Madras  
Presents**

**NPTEL  
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING**

**Pattern Recognition**

**Module 02**

**Lecture 06**

**Normal Distribution and  
Decision Boundaries I**

**Prof. Sukhendu Das  
Department of CSSE, IIT Madras**

Welcome back to the lecture on normal distribution and decision boundaries under the course titled pattern recognition. In the last class you would have heard about concepts of base decision rule which form the basis of the base classifier okay. And what are the three components in the expression for the base when you come toward the probability, posterior probability for a particular sample.

The three terms one of them is the class prior, evidence okay. Now when you actually want to use the base criteria for performing a pattern classification task, you need to actually compute the probability densities for samples, for a particular class. And for that you need certain models, one of such commonly used models, not only in the field of pattern recognition, but other many scientific and engineering disciplines is the normal distribution.

So we will start with details of normal distributions and see that how, using this normal distribution and base decision criteria we come up with certain decision boundaries for the task of classification, for a particular task.

(Refer Slide Time: 02:03)

## **The NORMAL DISTRIBUTION**

The normal (or Gaussian) distribution, is a very commonly used (occurring) function in the fields of probability theory, and has wide applications in the fields of:

- Pattern Recognition;
- Machine Learning;
- Artificial Neural Networks and Soft computing;
- Digital Signal (image, sound , video etc.) processing
- Vibrations, Graphics etc.



So let us look into the slide and find out what normal distribution is. So the normal or a Gaussian distribution is very commonly occurring function in the fields of probability theory, but it is also very wide applications in many other fields. Examples of course include, pattern recognition, machine learning which we are involved in this course. Artificial neural networks, soft computing, digital signal processing, other fields of vibrations, graphics, any sort of modeling which you need.

Normal distribution is a very commonly used function to model certain distribution. So we will first see the formula of the normal distribution, and then see certain properties of the distribution. And then we will proceed towards ways by which this distribution can be used for classification task, which involves certain distance measures and classify theory.

(Refer Slide Time: 03:15)

Its also called a BELL function/curve.

The formula for the normal distribution is:

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right]$$

The parameter  $\mu$  is called the mean or expectation (or median or mode) of the distribution,

The parameter  $\sigma$  is the standard deviation;  
and variance is thus  $\sigma^2$ .



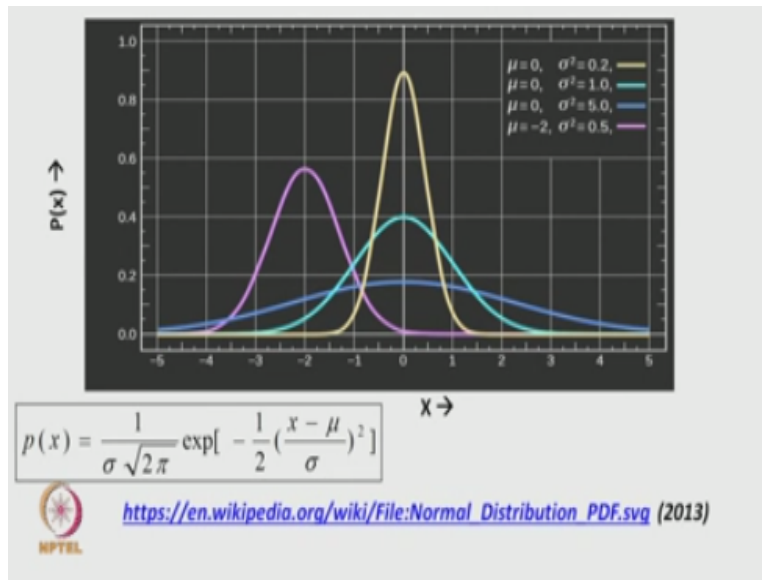
It is also called a BELL function, the Gaussian function is also called a BELL function or a BELL curve. And then formula of the normal distribution is given by this, it is a one-dimensional function, although this formula holds for multivariate random variables. So we will see that later. So for random variable X or for variable X, you see two terms, the denominator term on the first here is basically a normalizing term.

And the expression is usually given by this. You can see two parameters in the distribution, one is the  $\mu$  which is called the mean or expectation, some statisticians may say median or mode, but we will hence forth use the word mean of the distribution. There is another parameter  $\sigma$  which is actually more important in the normal distribution which is called the standard deviation, and square of the standard deviation is actually called the variance  $\sigma^2$ .

So I repeat again,  $\sigma$  is the standard deviation and  $\mu$  is the mean for that particular function. So it is a product of an exponential function of this feature divided by the normalizing term which is here. Now we will see some examples of this BELL function and the curve, some plots which will show you the importance of or the significance of these two parameters. Of course, mean  $\mu$  is very simple to interpret, it is like an average value.

So whatever be the average value of a set of random numbers or which is represented by X correct, the mean will represent that. How does  $\sigma$  that is the standard deviation control the nature of the BELL function or Gaussian function.

(Refer Slide Time: 05:37)



Let us look at this example in the slide. The reference of this particular plot is available from Wikipedia which is given below. So you can also have a look at this, but now observe there are actually four different curves in four different colors. And three of them have the same mean value 0. So the mean value 0 at this point is where the three curves in blue, sand and this yellowish curve is what you get grayish yellow.

One of the curves has a nonzero mean which is -2 as you see here in the top right, and that is given by the magenta curve. So that is how you can see this, how the mean value  $\mu$  helps you to position the curve, what is the effect of  $\sigma$ , while the values of  $\sigma^2$  are given four different values. And for the first three curves when the mean is 0, you can see that the curve in yellowish grey curve has the value of variance which is the least, because 0.2 is the variance, standard deviation will be root over that.

What about the one is sand color, the standard deviation of the variance is equal to 1, so that is the normal curve which you also define that when  $\sigma=1$ . And then what you also have is the one in blue where the variance is 5, which is the wider curve. You also have the nonzero mean Gaussian function when  $\mu=-2$ , the corresponding variance is 0.5 which you see here.

So if you compare these three curves having the mean at the center, you can see that if the variance is large the function has a large spread or width, or larger set of nonzero values, lesser the value of variance, more peaky or sharp or lesser is the extent or span or width of this

function. So what we just learnt now is, if you increase the variance of the, or standard deviation of the Gaussian function, you will have a larger span, the wave function will have a larger span or width.

If you have a small value of variance or standard deviation, you will have a much sharper or peaky nature of, almost if you take a very, very small value of  $\sigma$  that is the standard deviation, you will tend to actually create an impulse function, very narrow width that is for smaller values. For larger values, larger width, but lesser height.

What is the normalizing term being in the expression, if you go back to the slide look at the normalizing function the value of the function  $P$  of  $x$  or the Gaussian function is equal to this when this exponential value will be equal to 1 only under one condition this value of the exponent will be equal to 1 when will it be the value of the exponent will be = 1 look back the expression and think and tell when will with this term be equal to 1 when the value of  $x$  = the value of name.

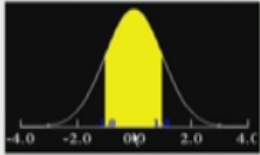
Okay so you can see from the curves back now that when the value of  $x$  touches the mean it the value of the function  $P(x)$  is also at it is maximum and then maximum value is dictated by this left hand side term here 1 by this is denominator where you have a  $\sigma$  that means for a larger value of  $\sigma$  which as a larger width you will have a lesser high for a smaller value of  $\sigma$  when you have a lesser span you will have a very large peak like an inverse.

What is actually happening is that the area under the curve is always same the normalizing term ensures that area under the curve is = 1 you can have a Gaussian function without the normalizing term the nature will be the same but the peak value always start at 1 it will not have that normalizing factor of having a area under the curve = 1 and hence the peak changing with the values only the width will change okay we will keep on looking at few more examples as we long.


(Refer Slide Time: 10:55)

**The 68 - 95 - 99.7% Rule:**  
**All normal density curves satisfy the following property**  
**which is often referred to as the Empirical Rule:**

- 68% of the observations fall within  
**1 standard deviation of the mean,**  
**that is, between  $(\mu - \sigma)$  and  $(\mu + \sigma)$**



The figure shows a normal distribution curve centered at 0. The x-axis is labeled with -4.0, -2.0, 0.0, 2.0, and 4.0. A yellow shaded region is shown between -1.0 and 1.0 on the x-axis, representing the area under the curve within one standard deviation of the mean.

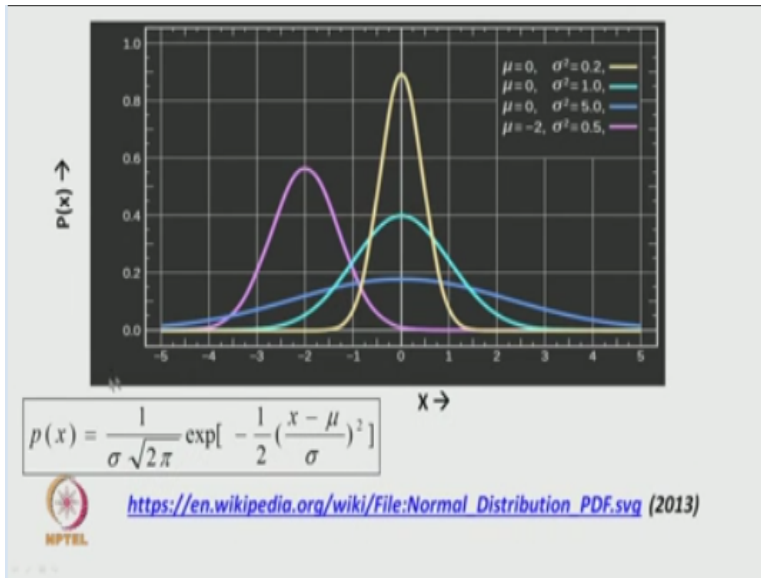
 NPTEL

Now this is a nice empirical rule which is sometimes causally called the 68- 95 and 99.7 or sometimes called 99.8 some books will use 99.7 or 8 rule in which a normal density curve or sometimes also called the Gaussian function okay we will interchangeably use hence forth in this lecture normal density or Gaussian function it satisfies the following property which is often referred to as the empirical rule.

What does that mean there are 3 parts to this whole it says that 68% of the observation that means you making density observations which is creating the density they fall with 1 standard deviation of the mean that is between  $\mu - \sigma$  so look at this curve of the plot of the Gaussian which is a colored in yellow from mean which is  $= 0$  in this case to  $-1$  and  $\mu + 1$  this images also you will get in certain websites if you start looking at properties or empirical rules of Gaussian function.

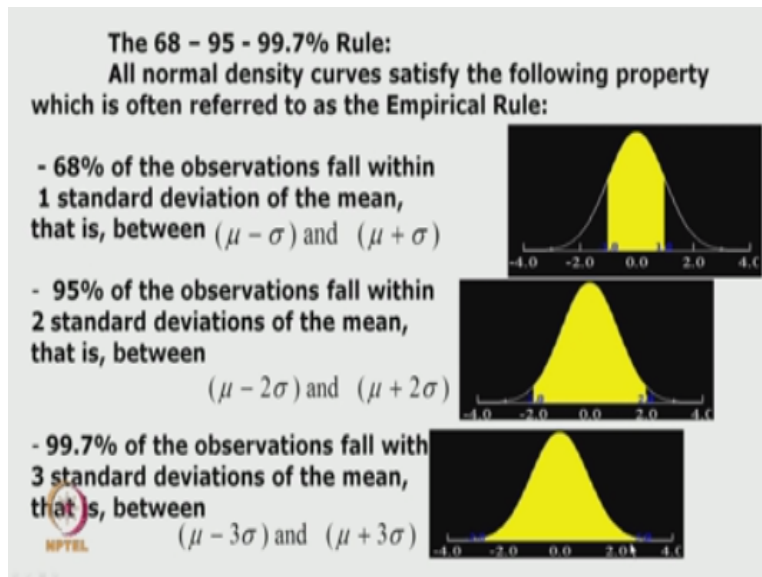
Okay let us go look back again so if you sum up the area of this yellow region from  $\sigma - 1$  which is here to  $\sigma + 1$ ,  $\sigma - 1$  to  $\sigma + 1$  the total area under the curve will be 68% of the total that means the total, total area of the normalized Gaussian curve actually this a Gaussian with  $\sigma = 1$  and how I will complete that in movement with another property. This is a Gaussian function with  $\sigma = 1$  did you have it in the previous slide  $\sigma = 1$ .

(Refer Slide Time: 12:43)



You look at this the same curve okay  $\sigma=1$  this is the one the blue color is been drawn here.

(Refer Slide Time: 12:55)



But we have shaded in yellow the area under the curve between  $\sigma - 1$  to  $\sigma + 1$  and it is 0.68 that means 68% total is = 1 what about something more if we go to some  $\sigma - 2$  sorry  $\mu - 2\sigma$  I repeat

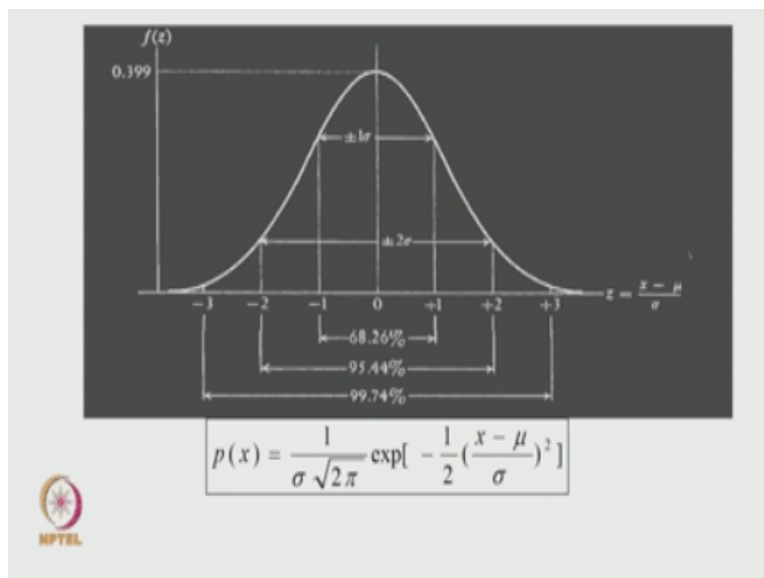
again if you start from  $\mu - 2\sigma$  since  $\sigma = 1$  we starting from -2 here you can see so -2 to +2 you look at this range here so you are starting from  $\mu - 2$  because  $\sigma = 1$  to  $\mu + 2$  you have 95% of the area curved with this range.

And if you still go one step further between  $\sigma - 3$  to +3 it seemed that you have all most curved the entire under the curve a negligible part is left beyond it is just 0.3 % so 99.7% of the observation of this area falls between  $\mu - 3\sigma$  to  $\mu + 3\sigma$  whatever be the value of  $\sigma$  this holds good the curve which you have seen are for  $\sigma = 1$  opr standard deviation = 1 only okay but this is valid for  $\sigma$  okay let us look back the curves I again I repeat.

So  $\mu \pm \sigma$  if you want to say it in a very gradual simple manner then that area is about 68 or 70% mean  $\pm 2\sigma$  is about 95% is and almost close to 100% is  $\mu \pm 3\sigma$  so that means if you taking observations from mean  $-3\sigma$  to  $\mu + 3\sigma$  you are actually almost capturing all the samples are all the observations much less than 1% is outside that range and that is why this range of  $6\sigma$  some books may taken even  $7\sigma$ .

So this is called  $6\sigma$ ,  $7\sigma$  rule empirical rule where you take the interval of  $\pm 3\sigma$  or  $\pm 3.5$  also some books we will find that this actually contains almost all the energy information or observations which you are trying to model using this Gaussian function.

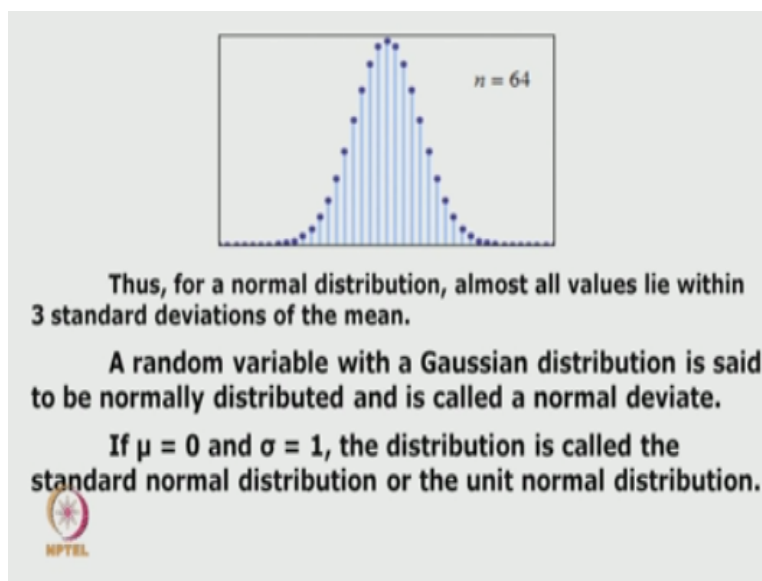
(Refer Slide Time: 15:19)





This is another curve from other source from other document the expression is given at the bottom and you can see the value 68, 95, 99.7 and  $9.8 \pm 3\sigma$  you want to go little bit beyond again  $\mu$  is consider to be 0 you look at the range  $\pm 1\sigma$  this range  $\pm 2\sigma$  and then  $\pm 3\sigma$  to so  $\sigma$  here okay it curves the entire range can you guess why this value of the peak is marked as closed to 0.4 0.399 but you can safely approximate it to 0.4 what could be the bases of that it is not a difficult answers you can think for a few more minutes and tell me based on last 5 minutes of whatever I have discussed and guess who what is that value of 0.4 representing.

Of course it is the peak of the Gaussian but what sort of the calculations is giving that is the simple question  $1/\sqrt{2\pi}$  that is correct because  $\sigma=1$  I did mention that this is the plot for standard deviation = 1 so I will leave it as an exercise for you to use your calculator to check the value of  $1/\sqrt{2\pi}$  we will get the value of 0.4 having very close 0.398, 2.4 is what you will get good. (Refer Slide Time: 16:53)



So what you have just observed now for a normal distribution almost all the values lie between or within 3 standard deviations of the mean the couple more statements to made before you move on to properties of Gaussian distribution a random variable with the Gaussian distribution is said to be normally distributed and is called a normal deviate if the mean = 0 which you see at the end and the standard deviation of variance = 1.

This distribution is called the standard normal distribution or unit normal distribution this is just form normal so somebody says what is a unit normal distribution or standard normal distribution

it is the Gaussian function with  $\sigma=1$  that is all okay and you can use the expression give in the last few slide to compute the curve okay you will see the animation now coming you in the slide in which actually it shows the Gaussian function with increasing values of points on the curve this is not actual observations made for the Gaussian we will come to that environment because this density function is basically probably.

So in the number of points are plotted more because this is an important thing to observe although you will not appreciate it right now because it is possible that in many applications of pattern recognition and signal processing you may have a very few observations sometimes to make. And often to may be that statically the number of values which are using for calculations either to compute a probability function or density function or may be an histogram let us say the number of points which you use to make the observations.

And the number of discrete values are the sample which you get from the density function it is better if the number of points are more then you have much more can see the value  $10 = 5$  look at the values just more then we saying the Gaussian function is not showing up in this smooth way as it is expected to okay this is just a nice animation to show you that a most examples of probability and statics we expect the values of the samples to be large number more larger to get.

(Refer Slide Time: 19:35)

**The normal distribution  $p(x)$ , with any mean  $\mu$  and any positive deviation  $\sigma$ , has the following properties:**

- **It is symmetric around the mean ( $\mu$ ) of the distribution**
- **It is unimodal: its first derivative is positive for  $x < \mu$ , negative for  $x > \mu$ , and zero only at  $x = \mu$ .**
- **It has two inflection points (where the second derivative of  $f$  is zero and changes sign), located one standard deviation away from the mean,  $x = \mu - \sigma$  and  $x = \mu + \sigma$ .**
- **It is log-concave.**
- **It is infinitely differentiable, indeed supersmooth of order 2.**



Let us look at some very nice important some more properties one important properties which we are seen now is that the effect of the standard deviation larger or wider okay and lesser height

smaller the value smaller the spread extent more peak it is okay, but these are imagine trust importance so the mean is  $\mu$  and the standard deviation  $\sigma$  is positive if you go back into the expression mathematically it is possible to actually have a negative value of  $\sigma$  mathematically possible it does not alter the value of the expression within the exponential because you have a square term.

But what will happen is if  $\sigma$  is negative the probability density becomes negative so it is basically a meaningless effect to choose a value of standard deviation which is negative okay deviations have a certain value has to be positive and density functions are usually positive and we will soon go add and use them as distance functions so they have to be all positive good, so do not worry about negative so we are any way has to be positive  $\sigma$  first fall it is symmetric around the mean.

We have seen that from the curves got the function is symmetric it is uni-modal it is first derivative is positive 1 the value of  $x$  is less than the mean and it is negative and it is more than in exactly 0 only at mean  $\mu$ , so the first derivative as a nice property connected to the secondary derivative is a very interesting property which says that the Gaussian function has two inflection points basically it is second order from located to one standard deviation away from the mean so at mean  $\mu - \sigma$  and  $+\sigma$  you will get two inflection points.

May be some features scope of analysis I usually the derivatives are the Gaussian but I leave this is an exercise for you to actuate the derivatives of the Gaussian first second even higher but at least stand in first and second order and just a good deal of application in many analysis of pattern recognition as well as image processing as well as image processing people let up edges with the help of derivatives of the Gaussian function which basically become edge operators okay so the property.

Which we only understand here is now that the first derivative is positive for negative values of  $x$  and negative for positive values of  $x$  0 only at the mean and it has two inflection points at these values it is log concave a function satisfies certain property to be a concave function and not getting into those details but you can find that out from certain concepts of algebra setting.

Concave function find that out if the log of that function is also concaved it is called a log concave function so this Gaussian function has that special property you may not use all of these

properties okay the last of the most important part why it is also popular is that it is infinitely differentiable look at the last sentence infinitely differentiable and it in deeds super smooth of order to okay infinitely differentiable that means we are only talking about first and second derivatives but I tell you that you can take even higher derivatives of the Gaussian.

As high as you can think and this gives scientist lot of many varying capability with the analytics with the help of Gaussian function if it is if you are able to model it that is the main reason of it is popularity number, the function stratifies certain criteria to be smooth again like the log concave of the concavity property of a function that is the function concave when it is smooth there are certain inequalities are criteria to be find that and there is a super smooth property of a curve of certain order b time is the order  $\beta$  is considered to be 2 the Gaussian function satisfies.

Certain properties curve is expected to be smooth it is better but in this case the last two hour important properties but they are probably would not be used in a great extent in our theories I think it is infinitely differentiable is 1 property sometimes high derivative of the Gaussian functions are used that the first and second derivative are very, very important.


(Refer Slide Time: 24:49)

**Also, the standard normal distribution  $p$  (with  $\mu = 0$  and  $\sigma = 1$ ) also has the following properties**

- Its first derivative  $p'(x)$  is:  $-x.p(x)$ .
- Its second derivative  $p''(x)$  is:  $(x^2 - 1).p(x)$
- More generally, its  $n$ -th derivative :

$p^{(n)}(x)$  is:  $(-1)^n H_n(x)p(x)$ ,

where,  $H_n$  is the Hermite polynomial of order  $n$ .



Yeah this is the expression of it is first derivative and we have taken mean to be 0 and  $\sigma = 1$  leave it as exercise for you to do the same when there is a non zero mean that means derive the expression of the derivative 1 and 2 derivative of the Gaussian function only mean is non zero and value of the center deviation is not equal to and you should be able to write it terms of the

Gaussian function like we have done it, that should be also possible if you look back so in this case mean is 0 and  $\sigma$  is = 1 but if it is not those parameters will come in the expression look at the second derivative of the function.

We can actually yourself from this see  $p(x)$  is always positive look at the second derivative of the expression here  $p(x)$  is always positive you know that from the function this component will be equal to 0 at only two values of  $x$  what are those two values I reap at if you look back into the expression this component will be equal to 0 at only two possible values  $x = +$  or  $- \sigma$  to be very precise because in this case  $\sigma = 1$  what are those let us go back to the previous slide is even at this point we just talking about.

Inflection points this is where the second derivative is 0 and it changes sign is sometimes called 0 cross in point also and it occurs at  $+$  and  $- \sigma$  here since  $\sigma = 1$  the value will so you can almost blindly close here all is and replace this by  $x^2 - \sigma^2$  is not it more general the  $n$ th derivative is given by this particular  $h$  of  $n$ th is given as the Hermite polynomial of order  $n$  and this is a common function used in interpolation in the field of comparative graph  $x$  okay not get into details by the Hermite polynomial of order  $n$  is a very common function used in many branches of science and engineering specifically car fitting in the field of interpolate graphics interpolation this taken.


(Refer Slide Time: 27:15)

**Normal Density:**  $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$

**Bivariate Normal Density:**

$$p(x, y) = \frac{e^{-\frac{1}{2(1-\rho_{xy}^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{2\rho_{xy}(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}}$$

$\mu$  - Mean;  $\sigma$  - S.D.;  $\rho_{xy}$  - Correlation Coefficient



So carrying the discussion on Gaussian or normal density function if you look back into the screen now the expression of one dimensional Gaussian function or normal density function is given. Let us look at the density function in two dimensions or what is called as the Bi-variate normal density function.

So you have two variables now instead of only x you have x and y, instead of just one standard deviation you have  $\sigma_x$  and  $\sigma_y$  that means standard deviation along x and y respectively you have the corresponding means as well  $\mu_x$  mean along x direction and  $\mu_y$  is for the y component. In addition you will also have a correlation coefficient so  $\mu$  stands for mean with the corresponding subscripts indicating the direction or component.

Standard deviation  $\sigma$  stands for standard deviation with its corresponding components and you have the correlation coefficient  $\rho_{x,y}$  which is the correlation coefficient between two variables x and y, okay. There is relationship between the correlation coefficient and the corresponding standard deviation and the joint covariance term between the two variables x and y we will have a look at that now.

(Refer Slide Time: 29:02)

**Covariance of x and y, is defined as:**

$$\sigma_{xy} = E[(x - \mu_x)(y - \mu_y)]$$

**Covariance indicates how much x and y vary together. The value depends on how much each variable tends to deviate from its mean, and also depends on the degree of association between x and y.**

**Correlation between x and y:**

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = E\left[\left(\frac{x - \mu_x}{\sigma_x}\right)\left(\frac{y - \mu_y}{\sigma_y}\right)\right]$$

**Property of correlation coefficient:**

$$-1 \leq \rho_{xy} \leq 1$$



The covariance of x and y is defined as the expectation of the two variables subtracted with their corresponding mean subtracted that means  $x - \mu_x$  represent the values of x centered around 0 the same with respect to y. The covariance indicates how much an x and y vary together we will see that with some examples after some time and the value of this covariance term depends on how much each variable tends to deviate from its mean as well as also how it depends on the degree of association between x and y.

So a larger relationship between x and y of course a mathematical relationship that means what type of relationship you may have well let us say if the value of x is rising will the value of y also rise or will it fall down or will it remain constant, uniform does not change. There may be many different possibilities and if it rises or falls does it do that steeply or gradually so these three different conditions along with their rate of change forms the value measuring the degree of association which is actually call the covariance term and it has a very strong relationship with the correlation coefficient which we saw inside the bi-variate normal density function, look back into the slide.

So the covariance indicates how much x and y vary together if you remember this it is quite sufficient with the time being it also depends on the degree of association between x and y. The correlation coefficient between x and y as the function of the covariates term  $\sigma_{xy}$  is given by this expression and in such a case you rewrite the covariance term using this.

So it is something like a normalized change okay you have the  $\sigma_x$  and  $\sigma_y$  is coming with the denominator and the correlation coefficient is actually a scalar quantity the value always lies between  $-$  and  $+1$  indicating the degree or relationship variance or association between  $x$  and  $y$ . This is an important formula which you remember that means you can actually rewrite covariance as correlation coefficient multiplied by the individual variances or standard deviation. Standard deviation we will precise, I repeat again covariance of  $x$  and  $y$  is correlation coefficient and multiplied by the individual standard deviations of  $x$  and  $y$ , okay.


(Refer Slide Time: 31:59)

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = E\left[\left(\frac{x - \mu_x}{\sigma_x}\right)\left(\frac{y - \mu_y}{\sigma_y}\right)\right]$$

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}}$$

The **correlation coefficient** can also be viewed as the cosine of the angle between the two vectors ( $\vec{x}$  and  $\vec{y}$ ) of samples drawn from the two random variables.

This method only works with centered data, i.e., data which have been shifted by the sample mean so as to have an average of zero.



This is another way by which you can write some books you will find that you will write the expression of correlation coefficient using expectations of  $x$  and  $y$  separately or using joint expression in this particular form you can use any one of these expression. This correlation coefficient is quite important because we would like to know how to variables depend on each other.

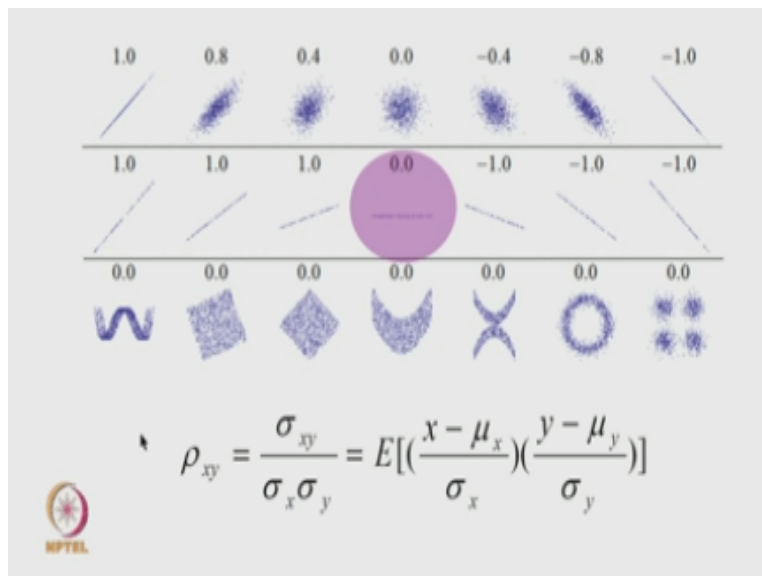
In our case in the field of pattern recognition we must remember that each of these variables are feature dimensions okay, so it is something like how one feature is related to the other one, okay are they joint as related or they heavily correlated or not, okay. The correlation coefficient and the covariance terms will be actually holding that information so you may need to estimate these parameters of correlation coefficient or covariance from the data itself and that is why we were looking at this particular formulas.



So again looking back to the slide what does correlation coefficient tell us? It is basically the cosine of the angle between the two vectors of course in three dimensional space but you can take D to be 2 when you are talking about only x and y of the samples drawn from two random variables.

So if we have two random variables x and y we were talking of just two vectors in two dimensional space and of course the data must be always be normalized or it is centered as it is called a shifted by the sample means so as to have an average of zero. This must be done for all analysis of classification and pattern recognition tasks.

(Refer Slide Time: 33:52)



This figure is a geometrical illustration of correlation coefficient values in 2D there are three rows of set of points or instants or samples drawn from a particular data and depending on the

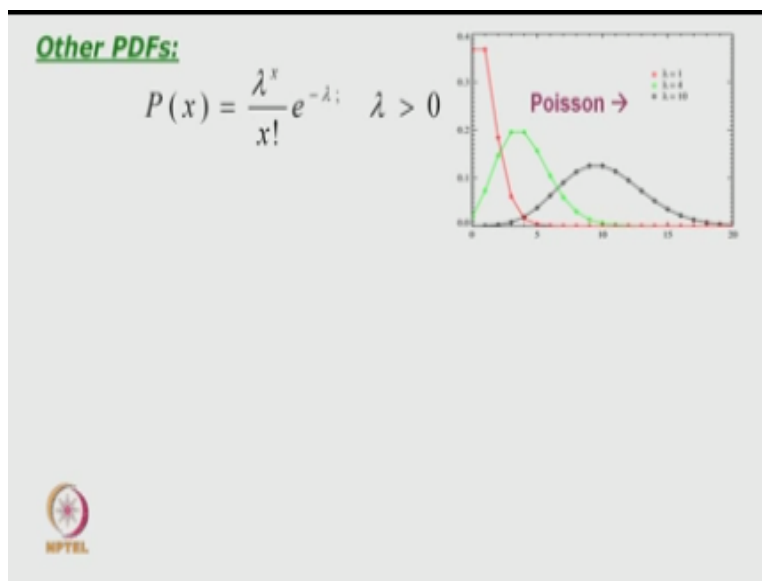
geometrical arrangement which you see the corresponding correlation coefficient which you have are given next to the figure.

Look at the last row as a typical example, the correlation coefficient is 0 what basically means that in some sense there is no relationship that means that there is no relationship between the two independent variables x and y, x and y are independent in some sense you can say. In the all the other cases you have a value of 1 here which basically means that this is a strong co-relationship between x and y no correlation again.

And a negative value of correlation that means when x is increasing it seems y is decreasing okay, you look at from the center if x is increasing in this direction the y is decreasing, this will give you an idea of why you have a negative value of co-relationship the same thing here. the value is not equal to -1 but still you can see that when x is increasing the y is decreasing and y so add.

So if x is decreasing y is increasing so there is a negative correlation shape here there is a positive, because both are either together increasing or together decreasing okay, the first gives an idea why you have positive values of correlation coefficient here on the left hand side and negative values here. So this gives you an idea about how the correlation coefficient may accept, now look at the value here which is 0.

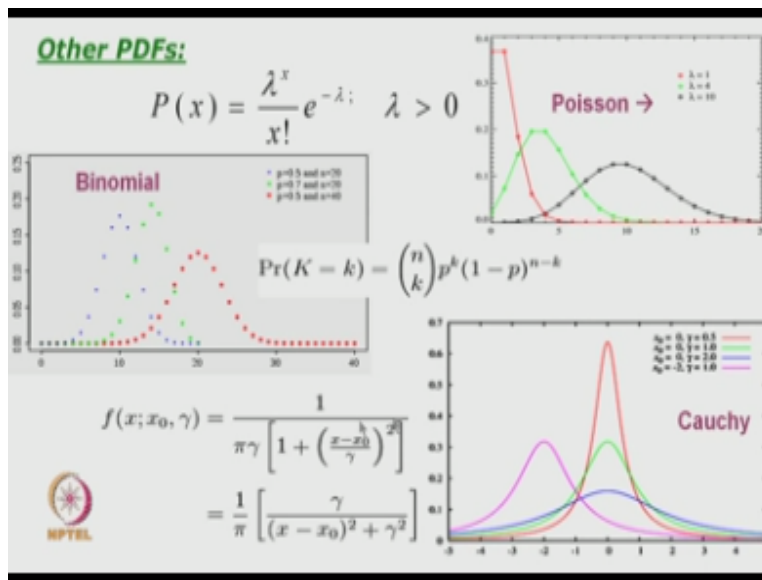
(Refer Slide Time: 35:48)



There are other examples of density functions which exist in the field of probability theory and statistics, but I talked well before hand as to why Gaussian is popular in the field of pattern recognition you know let me work signal processing other fields of mathematics and engineering as well the couple of important properties is this nice smoothness of the function is derivative existing up to and almost an infinite order and so on so forth which may not and it symitrisity many other nice properties which may not exist for other density functions.

They exist they can be used but you may not have the nice advantage of having mathematical manipulations or expressions done using no Poisson distribution functions will have just look at them a few of them some of them are popular but not that much to the extend as what moist mathematicians and scientist used in the field of signal image pattern recognition so on and so forth the Gaussian function.

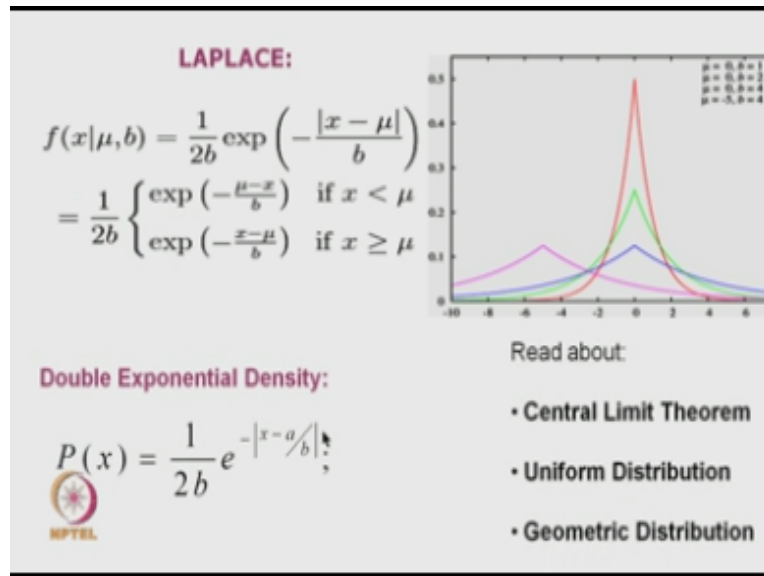
(Refer Slide Time: 36:57)



Well this is a Poisson distribution it is almost one sided and this is one parameter  $\lambda$  this is a binomial distribution it has two parameters n one is the value of n the other is the value of k then you have a Cauchy distribution where you have like similar to the Gaussian it is almost similar to the Gaussian except there is no exponential here but you can see the nature and look at the mean here the first three Are curves are for mean 0 and the last one is for the value -2 this is this curve.

And you have I think almost similar values is what we have used in our Gaussian function when we took the examples the Cauchy is very close to the Gaussian but its expression is different.

(Refer Slide Time: 37:55)



A Laplace function which is very peak in nature that means let us go back if you do not want this smoothness at the top for a particular density function, of course you could ask me a question now when do we need this function when do you need such a problem well let me tell you in spite of the Gaussian function being the most popular and the most commonly used ones it is possible that real data sometimes may not actually follow the Gaussian distribution in spite of the expectations of all theories scientist and engineers.

Specifically they are even use to model noise but unfortunately noise does not always follow the Gaussian distribution, any real left it I want to do for casting for the weather, stock market, elections, temperature okay see sight current whatever the case may be you can use also some model but real data usually does not follow always a very nice distribution and sometimes quite follow a from the Gaussian distribution.

Say it is good to have some other functions available at our disposal if we can use them and one such case which we are discussing now is that if you do not want the smoothness at the peak of the function let us say you want to peak in nature like this you can use a LAPLACE function

which has an exponential without a square term you can see this is similar to the Gaussian but first of all it has a parameter here but there is not square term which gives it this peak in nature here.

This is a double density expression which is expression wise similar without the square term again and I would encourage you to read other concepts if you have not gone through center limit theorem uniform distribution geometric distribution so on and so forth. So after we have studied a few examples of the properties of the Gaussian function in 1d and 2d let us now look in to the case when a Gaussian distribution is used to model a distribution in very high dimension.

What about the case when you have data in very high dimension you remember the lectures when we introduce the concepts of pattern recognition clustering classification that we have to extract a lot of feature from the data, the number of features which we extract from a particular signal could be a few tens to a few 100 to a 1000 in certain cases, so you may need to compute density distribution for features which are very large in dimension not only one and two.

So we need to have an expression now remember we had expressions of the Gaussian distribution in one d and 2 d I leave it is an exercise for you to write the expression of the Gaussian distribution in three dimension we have  $p(x)$  you had  $p(x,y)$  I leave it an exercise for you to write  $p(x,y,z)$  one simple extension will be instead of  $\mu$  you had  $\mu_x \mu_y$ , so you will have  $\mu_x \mu_y$  and  $\mu_z$  or you can write it as  $\mu_1 \mu_2 \mu_3$  three dimensions some books will follow 1 2 3 because you can write using this in this is rather than the  $x y z$ .

You will have the individual standard deviations or variances what are they? You had  $\sigma$ , then you had  $\sigma_x$  and  $\sigma_y$  you will now have  $\sigma_x \sigma_y \sigma_z$  or  $\sigma_1 \sigma_2 \sigma_3$  you did not have a correlation coefficient in one d you had one correlation coefficient in 2d, in 3d how many do you expect? I repeat again try to extrapolate the idea when you are doing in 1d you did not have any correlation because you did not have you cannot correlate if you just have one dimension data you have to correlate with something else correct.

So when you have 2 dimensions think about two dimensions  $x$  and  $y$  okay or direction one and two you had a  $\rho_{x,y}$  or a  $\rho_{1,2}$  if you think these are the two direction. Now you have three dimensions,  $x y$  and  $z$  you should be able to tell me how many correlations I can establish, three you have to take pairwise combinations that is what we did in 2d there only one option available okay.

So three of them now the question comes is there was three of those individual variances then three of those means expressions will be little bit complicated and it will get more and more complicated if you go to four or high dimension, so it is believed to have one expression which can handle in generally very large dimension  $d$  and then see if it can be generalized to one  $d$   $2d$  and  $3d$  let me tell you it will not be easy to write the expression in  $3d$  from  $2d$  although we have seen what are the extra terms and parameters required in 3 dimensional data for normal distribution correct okay.

But it would not be easy to extend that logic because you have to fit those correspondingly you can attempt to do that and the attempt will be more and more difficult for higher dimension, so let us have a very closed form nice compact expression in  $d$  dimension and see if we can write the two and the three as well which I leave it as an exercise for you, though we are looking at now if we look back in to the slide.

(Refer Slide Time: 44:03)

**Multi-variate Case:**  $X = [x_1 \ x_2 \ \dots \ x_d]^T$

Mean vector:  $\vec{\mu} = E(X) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix}$

**Covariance matrix (symmetric):**

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1d} & \sigma_{2d} & \dots & \sigma_d^2 \end{bmatrix}$$

NPTEL

You will have a multivariate case data that means in dimension you have random samples taken along is dimension is dimension could be a feature you know what are the features we are talked lot about them in our earlier classes. And for  $d$  dimension you have a  $d$  dimensional vector all the mean vector you see the notation use to show them that it is the vector okay, you can actually put the marker here also to indicate that this is also a vector, this is also a vector this is also a vector these are individuals scalar means along the corresponding directions.

So no problem in one d you will have just  $\mu_1$  in 2d you will have these two  $\mu_1 \mu_2$  or  $\mu_x \mu_y$  which we saw three dimensions you will have  $\mu_x \mu_y \mu_z$  or  $\mu_1 \mu_2 \mu_3$  now remains the most important point, the variances and co variances put together which is nicely we did not have to worry about this is one day because we just had one mean term and one variance. 2d we wrote an expression okay, 3d onwards for larger dimensions it becomes really complicated but there is close form expression and to do that we introduce matrix called the co variance matrix which will have all the terms which we are talking about.

It is usually denoted by  $\sigma$  but in some books we will find  $s$  as a symbol also look back into the expression, first of all you will see that this the symmetric manner both the matrix are symmetric either using left or right does not matter, this is the symmetric matrix as you can see here the individual variance are along the diagonal you can write it in this term as you like or in this whatever you feel.

Basically it is product of two individual standard deviation giving you the correspondences variances and you have a set of off back ground terms which is duplicated because it is symmetric okay, so if this is  $d \times d$  matrix can you tell me how many off term will be there? Not difficult if it is  $d \times d$  cross matrix off diagonal terms we will get totally, let us start with the total of diagonal terms that is very simple it should be  $d^2 - d/2$  because it is symmetric matrix.

$d^2 - d/2$  if  $d = 1$  dimension then the value will be 0 correct there is no term there  $d = 2$   $d^2 - d/2$  how much it will be there is just one of the term, we have that single  $\rho_{xy}$  just one term  $d = 3$  we talk about this sometime back, so you can see that this is the generic form of co variance matrix which can handle all these cases,  $d = 1, 2$  or very large dimension and this is the form which you must remember.

So it is the symmetric matrix and the diagonal terms have the certain significance the off diagonal terms certain other significance, the diagonal terms contain the individual standard deviation or the variance along the corresponding directions 1, 2, 3, so forth along the directions and the off diagonal terms are simply having as many possible corresponding, that means  $i^{\text{th}}$  term in that matrix will be giving you the co relation between  $i^{\text{th}}$  and the  $j^{\text{th}}$  directions and from that you can relate it to the co relation position  $\rho_{ij}$  or  $\sigma_{ij}$ .


(Refer Slide Time: 48:46)

**d-dimensional normal density is:**

$$p(X) = \frac{1}{\sqrt{\det(\Sigma)(2\pi)^d}} \exp\left[-\frac{(X - \bar{\mu})^T \Sigma^{-1} (X - \bar{\mu})}{2}\right]$$

$$= \frac{1}{\sqrt{\det(\Sigma)(2\pi)^d}} \exp\left[-\frac{1}{2} \sum_j (x_i - \mu_i) s_{ij} (x_j - \mu_j)\right]$$

where,  
 $s_{ij}$  is the i-j<sup>th</sup> component of  $\Sigma^{-1}$   
 (the inverse of covariance matrix  $\Sigma$ ).



So using that co variance matrix this is the d dimensional normal density that function is given by this expression I was telling you this is the co variance terms and you look you are actually using the inverse of the co variance matrix is given here, x and  $\mu$  are defined in the previous slide we will go back and have a look at it, so that is the data that is the mean vector and we already talked about  $x - \mu$  that means centre shifting it making the mean 0 this is what this will do and of course the normalizing factor here.

You can either use this expression in the form of a matrix or you can write it terms of this as well which is  $\sum$  of certain terms consisting of elements  $s_{ij}$  of the matrix which is the I j<sup>th</sup> component of the inverse of the co variance matrix, so the co variance matrix inverse as to be taken and then it has to used in this expression to compute, this is the d dimensional normal density functions and I would request you to almost attempt in due course of time few lectures to almost memorize this by heart.

As much as possible because it will be used thoroughly in our analysis in many places for clustering, classification, distance measures, and we will talk about more of that later in the next few classes as well okay, where you just need to remember that there is the normalizing term due to determinate of this co variance term the dimension d is also here  $2\pi$  and  $x - \mu$   $\sigma$  there is the T here okay.



You can actually remember this but this is the best thing to follow because hence forth we will keep on concentrating on different properties of this co variance matrix or it is inverse in various forms and see the net result of classification. The co variance matrix almost says about what the classification task you are trying to solve okay you might have heard somewhere in our earlier discussion I am going to hear a lot more in future about linear decision boundaries and non linear decision boundaries.

Class separate, distance between the clusters we talked about it in very beginning using an animation all those will get reflected in that co variance matrix, it will hold all the information of the data except the class means information is now here in this 2 terms but this co variance matrix will hold all the information about the way the data varies along each heavy direction and relationship between the two directions. So please remember this expression here as given by the.

(Refer Slide Time: 52:17)

$$p(X) = \frac{1}{\sqrt{\det(\Sigma)(2\pi)^d}} \exp\left[-\frac{(X - \mu)^T \Sigma^{-1} (X - \mu)}{2}\right]$$

$$= \frac{1}{\sqrt{\det(\Sigma)(2\pi)^d}} \exp\left[-\frac{1}{2} \sum_i (x_i - \mu_i) s_i (x_i - \mu_i)\right]$$

**Special case, d = 2; where X = (x y)<sup>T</sup>; Then:  $\vec{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$**

**and**  $\Sigma = \begin{pmatrix} \sigma_x^2 & \rho_{xy} \sigma_x \sigma_y \\ \rho_{xy} \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho_{xy} \sigma_x \sigma_y \\ \rho_{xy} \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix}$

**Can you now obtain this:**

$$p(x, y) = \frac{e^{-\frac{1}{2(1-\rho_{xy}^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{2\rho_{xy}(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}}{2\pi\sigma_x\sigma_y\sqrt{(1-\rho_{xy}^2)}}$$

It will repeated in the next slide as you can see here and special case where d = because this is where we talked about the bi variated normal density where the X is given here x,y corresponding mean vector also  $\mu_x, \mu_y$  and the co variance matrix will be at 2 x 2 because d = 2b

and as promise earlier you have the variance from the diagonal of terms what are these co variance terms.

So if the dimension  $d = 2$  you can see this  $d = 2$  and the  $\sqrt{\quad}$  will cancel out you can simplify this expression  $2 \times 2$  it is very easy to compute and what will happen here is the inverse of this matrix will it is also not difficult to compute  $2 \times 2$  matrix easy to compute. Using the  $\sigma$  substituting it here you get back this expression which we had a few slides back what was the expression the bi variated normal density with the co relation, co efficient.

And the corresponding to 2 different means I did tell you that this  $\sqrt{2\pi}$  will be available here the determinate of this should be sitting here and this is the exercise will show that and rest of it take the inverse substitute to it this is what you will get. Actually it is a normalizing term here also which will appear out of the determined of this projects because inverse will also have factor with terminal I leave this analytical derivation to you as a home task because this will help you to get used to this explosion help you to understand and also memories as much as possible a task of this expression of the normal.

(Refer Slide Time: 54:36)

$$p(\mathbf{X}) = \frac{1}{\sqrt{\det(\Sigma)(2\pi)^d}} \exp\left[-\frac{(\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})}{2}\right]$$

$$= \frac{1}{\sqrt{\det(\Sigma)(2\pi)^d}} \exp\left[-\frac{1}{2} \sum_i (x_i - \mu_i) s_i (x_i - \mu_i)\right]$$

**Special case,  $d = 2$ ; where  $\mathbf{X} = (x \ y)^T$ ; Then:  $\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$**   
**and**

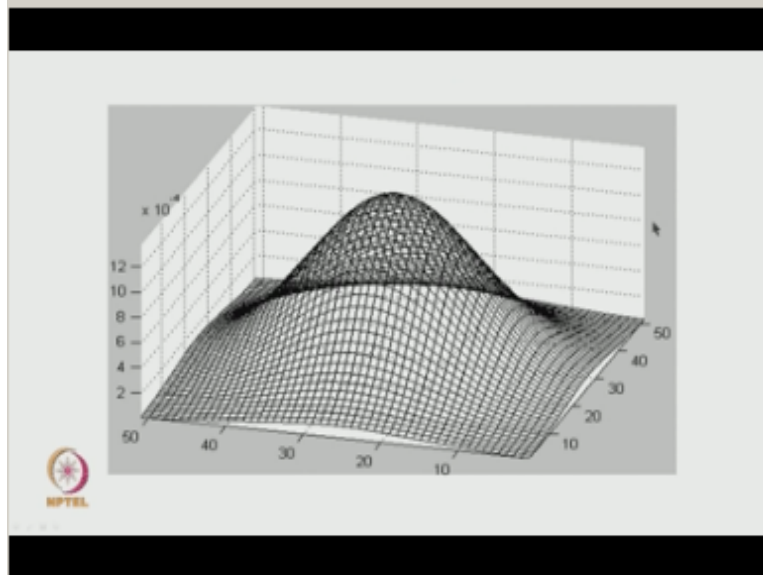
$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho_{xy} \sigma_x \sigma_y \\ \rho_{xy} \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix}$$

**Can you now obtain this:**

$$p(x, y) = \frac{e^{-\frac{1}{2(1-\rho_{xy}^2)} \left[ \frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho_{xy}(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right]}}{2\pi \sigma_x \sigma_y \sqrt{1-\rho_{xy}^2}}$$

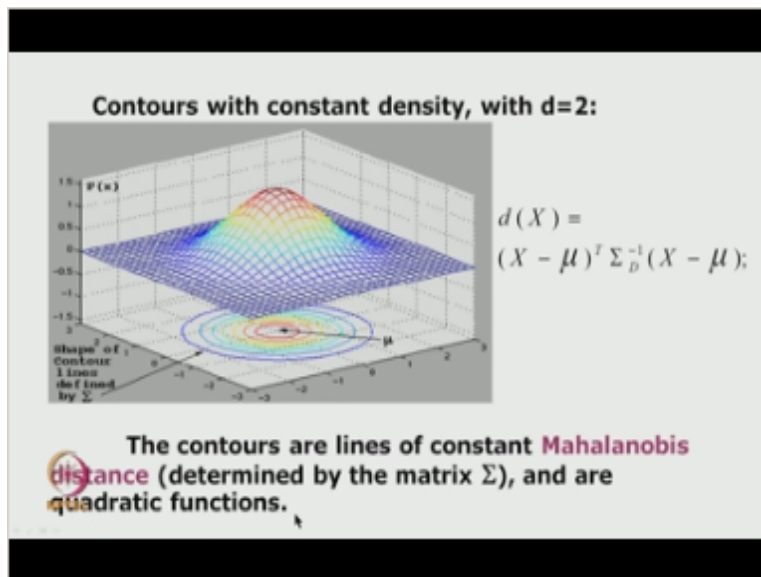
Three dimensional distributions for gossosons.

(Refer Slide Time: 54:39)



This is a picture it shows a wire frame diagram of two dimensional functions it is a very standard method of representing surfacing in the field of computer graphics so this is basically called mesh diagram of such a surfacing.

(Refer Slide Time: 54:58)



In 2d and what you see here are with that corresponding Gaussian function you are seeing at the bottom what are called as contours of lines which are at a certain distance as defined by the density function for the  $d=2$  that means we are talking about two dimensional case the Gaussian distribution function or the Gaussian density function and these circles are intersections of this Gaussian surface with horizontal plane of the value.

Given by the corresponding distance that means if I take this blue circular the outside which appears in the ellipse because of the projection basically it is a circle on that surface all the points in this surface are at equal distance is like a circle oversell it has to be equal distance from the center which is given like the new and that distance is some value which is defined by the covariance term.

And the density distribution function okay so closer this circular is to mean you are at a lesser distance or at a higher density you can think that if you have a Gaussian surface and find an intersection for a planner horizontal plane with that corresponding surface you will get a controller those conclude are shown below this surface at they are radically concentric circles with as you get out of the new away from you are talking about larger. And larger values of distance from the mean given by this particular expression now you see what is have done is just switch the logic this expression is same as the expression which we got in the previous satellites go back this expression what is have taken out is this particular it is

within the exponential and it can be easily taken it is a simple mathematical operation like as for example if you take a log of this expression.

Let us see equally in fact give after sometime next class you take a log of this expression you will get this expression now which you can actually is the key because it converts it actually contains the covariance matrix of dimensional its inverse and this is within the expansion I have now taken that term and say it is a distance  $d$  of a point  $x$  I repeat again it is distance  $d$  of a point  $x$  from the mean it is given you can suppress.

This term and use the rest of these two terms and you will also get a distance value which is typically what you will get is the I repeat again if you suppress the covariance term and take only these two and compute the distance you have this simple expression of distance which is called equilibrium distance very simple and that is a special case when the covariance matrix is matrix that is a special case covariance.

The individual variances are equal to 1 and the half diagonal term is 0 all the correlation coefficient all the coefficient are all 0 the diagonal terms which are sitting in the covariance matrix remember that expression which is asked you to almost memories as well this term it is an anti matrix now this is what we have so now simply we have new format distance function sorry I repeat again we have moved from a distribution to a distance relationship is very close a coherent term is there is both.

The distance from the mean is there in both only the normalizing term in the exponent is taken out to one because there in the expression of the Gaussians or normal distribution or density function in this case the distance is just we take that term within the exponents where it seems to give an exponent expression for a distance of a point that means what is the distance of this point to this point well you can measure it in two dimension  $x, y$  then  $u_1, u_2$  or  $v_x, v_y$ .

And compute that is the equivalent distance but if you want to take distributions of point into an account then you must use the expression of expanding the covariance functions as given in this slide this expression if you take it actually give you the correct distance in cooperating the density distribution and the distribution of points and these sort of the distance measure is actually called in the field of statics.

And estimation theory as well as pattern definition the Mahalanobis distance determined by the covariance matrix and these lines are usually quadratic functions well in this case they are acted

to elliptical or circular but they could be any other quadratic functions so we have just introduced the concept of distance from the normal distribution and we will now see how this distance placing a very important role in the job of classification.

Where you will now bring in concepts of what where did you have probability distribution for the classification talk so far we have disused one algorithm earlier that professor Moorthy that was the base rule for classification it had probability functions over density functions classifiers class conditions distribution if some of them are one of them are a Gaussian function if you put that concept now then we can derive distances out of rows to put inside the based decision rule so base decisions rule will.

Now become a pattern based on the distances instead of comparing probability we will compare distances we will consider this a nice correlation between the truth probability and distances because the expression of the Gaussian function allows you to do such operations in the already they can now repeat if we look back this is the expression we are talking about which contains the covariance matrix it was the probability distribution function.

And it is now taken as the distance now we will see using as a distance how it can be appropriated in the classification function that will give us the distance it will give a decision rules and it will give what are called decision boundary decision regions based on these certain discrete functions and you may get linear boundaries sometimes linear analysis next round of discussion we will flow these and we will go towards linear a non linear decision boundary as well as discreet analysis which we are only lead as direct analysis we will stop here.

### **Online Video Editing /Post Production**

K.R.Mahendra Babu

Soju Francis

S.Pradeepa

### **Camera**

Selvam

Robert Joseph

Karthikeyan

Ram Kumar

Ramganes

Sathiaraj

### **Studio Assistants**

Krishankumar  
Linuselman  
Saranraj

**Animations**

Anushree Santhosh  
Pradeep Valan .S.L

**NPTEL Web & Faculty Assistance Team**

Allen Jacob Dinesh  
Bharathi Balaji  
Deepa Venkatraman  
Dianis Bertin  
Gayathri  
Gurumoorthi  
Jason Prasad  
Jayanthi  
Kamala Ramakrishnan  
Lakshmi Priya  
Malarvizhi  
Manikandasivam  
Mohana Sundari  
Muthu Kumaran  
Naveen Kumar  
Palani  
Salomi  
Senthil  
Sridharan  
Suriyakumari

**Administrative Assistant**

Janakiraman.K.S

**Video Producers**

K.R. Ravindranath  
Kannan Krishnamurthy

**IIT Madras Production**

Funded By  
Department of Higher Education  
Ministry of Human Resource Development

Government of India

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved